Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Speech Communication 51 (2009) 83–96

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Feature compensation in the cepstral domain employing model combination

Wooil Kim, John H.L. Hansen [*]

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA

Received 11 May 2007; received in revised form 29 May 2008; accepted 16 June 2008

Abstract

In this paper, we present an effective cepstral feature compensation scheme which leverages knowledge of the speech model in order to achieve robust speech recognition. In the proposed scheme, the requirement for a prior noisy speech database in off-line training is eliminated by employing parallel model combination for the noise-corrupted speech model. Gaussian mixture models of clean speech and noise are used for the model combination. The adaptation of the noisy speech model is possible only by updating the noise model. This method has the advantage of reduced computational expenses and improved accuracy for model estimation since it is applied in the cepstral domain. In order to cope with time-varying background noise, a novel interpolation method of multiple models is employed. By sequentially calculating the posterior probability of each environmental model, the compensation procedure can be applied on a frame-by-frame basis. In order to reduce the computational expense due to the multiple-model method, a technique of sharing similar Gaussian components is proposed. Acoustically similar components across an inventory of environmental models are selected by the proposed sub-optimal algorithm which employs the Kullback–Leibler similarity distance. The combined hybrid model, which consists of the selected Gaussian components is used for noisy speech model sharing. The performance is examined using Aurora2 and speech data for an in-vehicle environment. The proposed feature compensation algorithm is compared with standard methods in the field (e.g., CMN, spectral subtraction, RATZ). The experimental results demonstrate that the proposed feature compensation schemes are very effective in realizing robust speech recognition in adverse noisy environments. The proposed model combination-based feature compensation method is superior to existing model-based feature compensation methods. Of particular interest is that the proposed method shows up to an 11.59% relative WER reduction compared to the ETSI AFE front-end method. The multi-model approach is effective at coping with changing noise conditions for input speech, producing comparable performance to the matched model condition. Applying the mixture sharing method brings a significant reduction in computational overhead, while maintaining recognition performance at a reasonable level with near real-time operation.

Keywords: Speech recognition; Feature compensation; Model combination; Multiple models; Mixture sharing

1. Introduction

The mismatch between training and operating environments is a significant factor that degrades the performance of speech recognition systems. Additive background noise, microphone mismatch, and channel distortion are typical sources of such performance degradation. Bridging the environmental mismatch gap for train/test material is one of the most essential issues in effectively addressing real-world applications using speech recognition technology

[*] Corresponding author. Address: Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA. Tel.: +1 972 883 2910; fax: +1 972 883 2710.

E-mail address: john.hansen@utdallas.edu (J.H.L. Hansen).

URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.06.004

and extensive research efforts from many groups have driven to realize this goal. Bridging the train/test environmental noise gap also requires the ability to detect, characterize and track environmental noise (e.g., *Environmental Sniffing* by Akbacak and Hansen, 2007) as well as understanding the challenges that exist in multiple simulation types of noise (e.g., *SoundScape* by Schulte-Fortlcamp et al., 2007).

The algorithms used to minimize the environmental mismatch can be generally categorized into two groups. One general class of algorithms focuses on migrating the input test data to be closer to the original training condition by compensating the speech signal or extracted features. Alternatively, the second category would concentrate on transforming the prior trained acoustic model to be closer to the test speech acoustics. Speech enhancement and feature processing such as Cepstral Mean Normalization (CMN) are examples of how to bring the input operating environment closer to the original training environment by suppressing noise or channel in the speech signal or extracted feature components (Boll, 1979; Ephraim and Malah, 1984; Lee, 1989; Hansen and Clements, 1991; Singh et al., 2002; Hansen and Arslan, 1995; Raj and Stern, 2005). Methods belonging to the second category are not directed at removing noise components, but generating a speech model which matches better the noisy environment during the training or decoding steps. The Maximum A Posteriori (MAP) (Lee et al., 1991; Gauvain and Lee, 1994) and Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) adaptation techniques are methods employed to re-estimate the acoustic models for an improved match to the test environment using available data. Parallel Model Combination (PMC), originally developed by Varga and Moore (1990) and refined later by Gales and Young (1996) generates a noise-corrupted Hidden Markov Model (HMM) by combining separate speech and noise HMM models.

This study focuses on developing an effective front-end feature compensation method to reduce the impact of additive background noise for robust speech recognition. A number of methods have been developed for front-end feature enhancement/compensation to improve speech recognition. An early approach for cepstral compensation was MCE-ACC by Hansen (1994) which employed adaptive cepstral compensation over voiced, transitional, and unvoiced detected segments over time. The motivation was to suppress spectral variability due to stress and emotion. Morphological constrained speech enhancement was employed to suppress noise. Other cepstral feature compensation methods later developed for speaker variability focused on fixed and adaptive cepstral compensation (Hansen and Arslan, 1995; Hansen, 1996) as well as neural network based methods (Hansen and Womack, 1996; Womack and Hansen, 1996). Various forms of cepstral mean normalization (CMN) have also evolved to address channel and microphone mismatch (Acero, 1993; Moreno, 1996). A second area for improving ASR is to use model-based feature compensation where models are developed for the speech signal typically using Gaussian mixture model (GMM). Transitions into noisy environments are characterized so that a clean speech feature response can be obtained using the transition compensation.

The model-based feature compensation methods can be classified into several categories according to how the noise-corrupted speech model is estimated. The first category is a data-driven method such as Multivariate Gaussian-Based Cepstral Normalization (RATZ), Stereo-based Piecewise Linear Compensation for Environments (SPLICE), and others (Moreno, 1996; Moreno et al., 1998; Droppo et al., 2001; Morales et al., 2006). Most of these methods require a noise-corrupted speech database to train the noisy speech model where the database is assumed to have acoustic characteristics identical to test conditions. In general, the training of the speech model is accomplished off-line, and therefore implementation is very efficient with limited required computational resources. However, when test conditions change from the training conditions, performance drastically decreases. An alternative category employs online estimation of the noisy speech model. Vector Taylor Series (VTS) (Moreno, 1996) and Interacting Multiple Model (IMM) (Kim, 2002) algorithms are representative examples for this category, which have the advantage of reflecting the current noisy condition by estimating the noise components from the incoming speech. A disadvantage is that the estimation procedure requires considerable computational resources. In particular for VTS, it has been shown that the performance does not outperform other model-based methods in our subsequent experiments. Feature compensation methods based on model combination can be considered as a third category. In these methods, the noisy speech model is estimated by combining the clean speech and noise model, which was originally proposed for HMM adaptation (Hansen, 1996; Womack and Hansen, 1999; Westphal and Waibel, 2001; Segura et al., 2001; Sasou et al., 2003; Kim et al., 2003; Stouten et al., 2004). The noise model can be obtained by off-line training or estimation from incoming speech. Most existing methods of the last two categories are applied in the log-spectral domain which have a larger number of coefficients than the cepstral domain. In addition, the methods often assume diagonal covariance matrices of the speech distribution, although the log-spectral coefficients are more highly correlated than the cepstral coefficients. These aspects not only increase the computational expenses, they also degrade the accuracy in estimation of the speech model.

In this study, we present feature compensation schemes employing model combination for noise-corrupted speech, which are applied in the cepstral domain (Kim et al., 2003, 2004). By using model combination, the proposed scheme eliminates the prior training which requires a noise-corrupted speech database, which is an absolute requirement in conventional data-driven methods. Independent access to the noise model makes adaptation in the non-speech interval possible. The advantages of the proposed method will be addressed in terms of model accuracy as well as efficiency applied in the cepstral domain. The novel interpolation method employing multiple environmental noise models is developed to address time-varying noise conditions. In order to reduce the computational expenses due to the combination of multiple models, a technique of mixture sharing is also presented.

The paper is organized as follows: the speech model-based feature compensation scheme is first reviewed and relevant issues are identified in Section 2. The proposed feature compensation method is described in Section 3. The multiple model approaches are presented in Sections 4 and 5. The representative experimental procedures and results are presented and discussed in Section 6. Finally, in Section 7, concluding remarks and a discussion of future work is presented.

2. Feature compensation method using GMM of speech distribution

Feature compensation employing a speech model has been considered by Acero (1993), and afterwards, Moreno designed a data-driven method which motivated similar schemes (Moreno, 1996; Moreno et al., 1998). In most speech model-based feature compensation methods, a statistical transformation of the clean speech's distribution under noisy conditions is estimated from the noisy speech, and then the noisy speech input is reconstructed using the estimated statistical variation. The speech model is generally estimated using a Gaussian mixture model (GMM). Each method has its own mathematical assumption for the relationship between the clean speech GMM and the noisy speech GMM. Some methods estimate the transformation by training the speech database off-line and others utilize incoming speech inputs.

In the general speech model-based feature compensation methods, the distribution of the clean speech feature $\mathbf{x}$ is represented with a Gaussian Mixture Model (GMM) consisting of K components as follows:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}), \tag{1}$$

where the speech feature $\mathbf{x}$ can be the cepstrum or log-spectrum depending on which method is employed. It is assumed that the noisy environment degrades by moving the means and the covariance matrices of the clean speech model of Eq. (1). Therefore, the distribution of the noisy speech $\mathbf{y}$ can be expressed as,

$$p(\mathbf{y}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}), \tag{2}$$

$$\boldsymbol{\mu}_{\mathbf{y},k} = \mathbf{f}(\mathbf{y}, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}), \quad \boldsymbol{\Sigma}_{\mathbf{y},k} = \mathbf{g}(\mathbf{y}, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}). \tag{3}$$

The functions $\mathbf{f}$ and $\mathbf{g}$ in Eq. (3) are based on assumptions on the transformation of the mean and covariance which are different for each method. MCE-ACC (Hansen, 1994) assumes a constant bias transform for the mean in the cepstral domain, while RATZ (Moreno et al., 1998) and SPLICE (Droppo et al., 2001) assume that the constant bias transform is for the mean and covariance. The VTS (Moreno et al., 1998) and IMM (Kim, 2002) methods employ a linear approximation of the relationship between the model parameters of clean speech and noisy speech in the log-spectral domain. Extensions to MCE-ACC (Hansen, 1994) include other methods that utilize model combination techniques (Hansen, 1996; Westphal and Waibel, 2001; Kim et al., 2003). Based on these assumptions, the mean and covariance of the noise-corrupted speech model of Eq. (3) are estimated from either the incoming noisy speech or a database constructed under an environment which is identical to the testing condition. In order to reconstruct the clean speech features from the noisy input feature vectors, the Minimum Mean Squared Error (MMSE) estimator is generally employed as follows (Ephraim and Malah, 1984):

$$\hat{\mathbf{x}}_{\text{MMSE}} = E\{\mathbf{x}|\mathbf{y}\} = \sum_{k=1}^{K} p(k|\mathbf{y}) E\{\mathbf{x}|k, \mathbf{y}\}. \tag{4}$$

The posterior probability $p(k|\mathbf{y})$ in Eq. (4) is given by

$$p(k|\mathbf{y}) = \frac{\omega_k p(\mathbf{y}|k)}{\sum_{k=1}^{K} \omega_k p(\mathbf{y}|k)}, \tag{5}$$

where $p(\mathbf{y}|k) = p(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k})$.

Data-driven methods such as MCE-ACC, RATZ and SPLICE have the advantage of being simple and fast computational procedures, however, they require off-line training using a prior degraded speech database. In addition, the performance of RATZ and SPLICE is drastically degraded when the testing condition does not match the training environment. VTS and other similar methods estimate the noise components from the incoming speech adaptively without requiring a training procedure. However, these methods require considerable computation in order to accomplish the iterative Expectation Maximization (EM) algorithm to estimate the noise components. Additionally, since they are applied to the log-spectral domain, which generally has higher dimension than the cepstral domain, the computational requirements become increasingly complex.

These previously developed feature compensation methods based on model combination are accomplished in the log-spectral domain, leading to an increase in computational expenses. The fact that most of these employ a *log-add* method for model combination to estimate the noise-corrupted speech model implies that they do not guarantee significant performance improvement because only the mean parameters of noisy speech model are estimated.

For feature compensation methods applied in the log-spectral domain, the speech models are estimated using covariance matrices which only have diagonal components. However, the log-spectral coefficients are more highly correlated with each other compared to the cepstral coefficients which are obtained through the Discrete Cosine Transform (DCT). Therefore, the presentation of the GMM for speech using a diagonal covariance in the log-spectral domain, which many current feature compensation methods employ, is a drawback which limits performance gain.

3. Feature compensation employing model combination

In this section, a novel feature compensation method is proposed in the cepstral domain which is based on a combination of Gaussian mixture models. First, we review the model combination method employed for generating the noise-corrupted speech model. Next, the details on how the model combination method is incorporated into the proposed scheme will be described.

3.1. Review of parallel model combination

Parallel model combination (PMC), first developed by Varga and Moore (1990) and later refined by Gales and Young (1996), assumes that a recognition system exhibits optimal performance when the training and test conditions are identical and the clean speech model is transformed into the noise-corrupted speech model to approximate the actual noisy environment. In order to generate the noise-corrupted speech model, the clean speech model and noise model are used independently. PMC is known to have many advantages, assuming the prior noise model and present noise environment have similar spectral and correlation properties.

Combining the speech and noise models is accomplished using the following *mismatch function*:

$$Y_i^{\{\log\}}(\tau) = \mathscr{F}\left(X_i^{\{\log\}}(\tau), N_i^{\{\log\}}(\tau)\right) = \log\left(\exp\left(X_i^{\{\log\}}(\tau)\right) + g\cdot\exp\left(N_i^{\{\log\}}(\tau)\right)\right). \tag{6}$$

Here $X_i^{\{\log\}}(\tau), N_i^{\{\log\}}(\tau)$, $Y_i^{\{\log\}}(\tau)$, and g denote the ith element of the clean speech, noise, noise-corrupted speech, and the gain respectively in the log-spectral domain. The random variables whose probability distribution is Gaussian in the log-spectral domain have a log-normal distribution in the linear spectral domain due to the exponential transform. In the *log-normal approximation* method, it is assumed that the addition of two log-normal distributions also results in a log-normal formulation. The mean and covariance of the corrupted speech are thereby computed by Eq. (7) based on this assumption and the mismatch function in Eq. (6)

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{y}}^{\{\text{lin}\}} &= \boldsymbol{\mu}_{\mathbf{x}}^{\{\text{lin}\}} + g\boldsymbol{\mu}_{\mathbf{n}}^{\{\text{lin}\}},\\ \boldsymbol{\Sigma}_{\mathbf{y}}^{\{\text{lin}\}} &= \boldsymbol{\Sigma}_{\mathbf{x}}^{\{\text{lin}\}} + g^2\boldsymbol{\Sigma}_{\mathbf{n}}^{\{\text{lin}\}}.\end{aligned} \tag{7}$$

Here $\boldsymbol{\mu}_{\mathbf{y}}^{\{\text{lin}\}}, \boldsymbol{\mu}_{\mathbf{x}}^{\{\text{lin}\}}$, and $\boldsymbol{\mu}_{\mathbf{n}}^{\{\text{lin}\}}$ refer to the mean vectors of the corrupted speech, clean speech and noise, respectively and $\boldsymbol{\Sigma}_{\mathbf{y}}^{\{\text{lin}\}}, \boldsymbol{\Sigma}_{\mathbf{x}}^{\{\text{lin}\}}$, and $\boldsymbol{\Sigma}_{\mathbf{n}}^{\{\text{lin}\}}$ denotes their corresponding covariance matrices of the log-normal distributions in the linear spectral domain. The mean and covariance of the linear spectrum with a log-normal distribution are obtained from the mean and covariance of the log-spectrum using the following equations:

$$\begin{aligned}\mu_i^{\{\text{lin}\}} &= \exp\left(\mu_i^{\{\log\}} + \Sigma_{ii}^{\{\log\}}\Big/2\right),\\ \Sigma_{ij}^{\{\text{lin}\}} &= \mu_i^{\{\text{lin}\}}\mu_j^{\{\text{lin}\}}\left[\exp\left(\Sigma_{ij}^{\{\log\}}\right) - 1\right].\end{aligned} \tag{8}$$

Finally, the mean and covariance of the corrupted speech in the linear spectral domain obtained from Eq. (7) must be converted back to the log-spectrum. The mean and covariance of the corrupted speech's log-spectrum are approximately calculated from the parameters estimated in the linear spectral domain using the following equations:

$$\begin{aligned}\mu_i^{\{\log\}} &\approx \log\left(\mu_i^{\{\text{lin}\}}\right) - \frac{1}{2}\log\left(\frac{\Sigma_{ii}^{\{\text{lin}\}}}{(\mu_i^{\{\text{lin}\}})^2} + 1\right),\\ \Sigma_{ij}^{\{\log\}} &\approx \log\left(\frac{\Sigma_{ij}^{\{\text{lin}\}}}{\mu_i^{\{\text{lin}\}}\mu_j^{\{\text{lin}\}}} + 1\right).\end{aligned} \tag{9}$$

3.2. Feature compensation via parallel combined Gaussian mixture model

In the proposed feature compensation method, the parallel model combination described in the previous section is employed to generate the GMM of the noise-corrupted speech (Kim et al., 2003). The proposed method is referred to as feature compensation based on Parallel Combined Gaussian Mixture Model (PCGMM) method. The proposed algorithm is based on the statistical distribution of speech features in the cepstral domain. The relationship between the cepstral feature vectors of clean speech **x**, additive noise **n** and noise-corrupted speech **y**, is presented as follows:

$$\mathbf{y} = \mathbf{x} + \mathbf{C}\log\left(\mathbf{1} + \exp\left(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{x})\right)\right) = \mathbf{x} + \mathbf{g}(\mathbf{x}, \mathbf{n}), \tag{10}$$

where **C** and $\mathbf{C}^{-1}$ denotes the DCT and its inverse transform, respectively. Based on Eq. (10), the relationship between the means of the clean and noisy speech can be derived as shown in Eq. (11). It is assumed that there is a constant bias transformation of the mean parameters of the clean speech model in the cepstral domain under the additive noisy environment, which is the assumption taken by other data-driven methods (Hansen, 1994; Moreno, 1996),

$$\boldsymbol{\mu}_{\mathbf{y}} = E\{\mathbf{x}\} + E\{\mathbf{g}(\mathbf{x}, \mathbf{n})\} = \boldsymbol{\mu}_{\mathbf{x}} + \mathbf{r}. \tag{11}$$

The relationship in Eq. (11) can be applied to each Gaussian component which composes the GMM of speech as follows:

$$\boldsymbol{\mu}_{\mathbf{y},k} = \boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{r}_k. \tag{12}$$

The bias terms $\mathbf{r}_k$ are used for reconstruction of the speech features. These values can be estimated with Eq. (12), once the mean parameters of the clean speech model and corresponding noise-corrupted speech model are obtained.

The clean speech model in the cepstral domain is estimated as a GMM through training on the clean speech database as shown in Eq. (1). The noise model is estimated as a single Gaussian model using the silence duration of the incoming speech or noise samples off-line. The noise-corrupted speech model is then obtained using the log-normal approximation method from Section 3.1. In order to combine the clean speech and noise models, it is required to convert the model parameters from the cepstral domain to the log-spectral domain. The mean and covariance of the cepstral domain are transformed to those of the log-spectral domain using an inverse DCT,

$$\begin{aligned}\boldsymbol{\mu}^{\{\log\}} &= \mathbf{C}^{-1}\boldsymbol{\mu},\\ \boldsymbol{\Sigma}^{\{\log\}} &= \mathbf{C}^{-1}\boldsymbol{\Sigma}(\mathbf{C}^{-1})^{\mathrm{T}}.\end{aligned} \tag{13}$$

In this study, the row and column of the DCT matrix $\mathbf{C}$ will have the same size as the number of log-spectral coefficients. Since the number of log-spectral coefficients is larger than the cepstrum, the mean and variance of the cepstrum in Eq. (13) will have additional padded zero values. After both models for clean speech and noise are converted into the log-spectral domain by Eq. (13), the model parameters of the noisy speech distribution can be estimated using the model combination procedure. Finally, the parameters of the noisy speech model must be returned to the cepstral domain via the DCT transform, which is the inverse process of Eq. (13). Now, the GMM of the noise-corrupted speech $\{\omega_k, \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\}$ is obtained in the cepstral domain and the constant bias term $\mathbf{r}_k$ of each component is estimated with Eq. (12). The MMSE equation for reconstruction of the clean speech in Eq. (4) is approximated with Eq. (14) in a manner similar to the method which was also used in (Moreno, 1996),

$$\begin{aligned}\hat{\mathbf{x}}_{\mathrm{MMSE}} &= \int_{\mathcal{X}} \mathbf{x}p(\mathbf{x}|\mathbf{y})\,d\mathbf{x} = \int_{\mathcal{X}} (\mathbf{y} - \mathbf{g}(\mathbf{x},\mathbf{n}))p(\mathbf{x}|\mathbf{y})\,d\mathbf{x}\\ &\cong \mathbf{y} - \sum_{k=1}^{K} \mathbf{r}_k p(k|\mathbf{y}).\end{aligned} \tag{14}$$

In Eq. (14), the variation of the cepstral feature represented by the function $\mathbf{g}(\mathbf{x},\mathbf{n})$ is replaced with the constant bias term $\mathbf{r}_k$ that depends on the Gaussian component index. Here, $p(k|\mathbf{y})$ can be calculated with Eq. (5), where the parameters of the noisy speech GMM $\{\omega_k, \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\}$ are obtained via model combination. Fig. 1 presents the resulting block diagram of the PCGMM-based approach as described here.

At this point, the distinguishing properties of the proposed method are considered, and compared with prior techniques. First, the new method does not require an additional training procedure using a noise-corrupted speech database. After obtaining the estimated noise model from the available noise samples, the distribution model of the noise-corrupted speech can be generated via the proposed model combination procedure. This results in a compensation method without the need of prior training data as seen in existing data-driven methods such as MCE-ACC (Hansen, 1994), RATZ (Moreno, 1996), SPLICE (Droppo et al., 2001) and others.

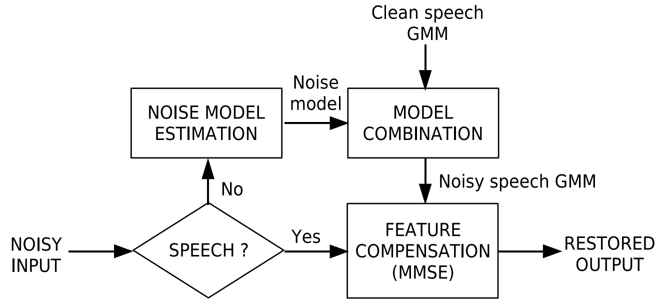

Fig. 1. Block diagram of the PCGMM-based feature compensation method.

The proposed PCGMM-based method employs a simple model combination procedure using noise model which is generally estimated as a single Gaussian model. Therefore, the proposed method does not require considerable computational expenses compared to VTS and other methods which estimate noise components by using an iterative EM process (Moreno, 1996; Kim, 2002) or utilize HMM update/decoding (Sasou et al., 2004).

In our proposed method, estimation of the GMMs for clean speech, noise, and noisy speech as well as the reconstruction procedure are accomplished all in the cepstral domain. The vector size of the cepstral coefficients is generally smaller than that of log-spectral coefficients, therefore, the PCGMM method has the explicit advantage of a lower dimensional space (e.g., reduced computation and storage) compared to other methods which operate in the log-spectral domain (Moreno, 1996; Segura et al., 2001; Sasou et al., 2003). In particular, the cepstral coefficients are less correlated with each other compared to the same coefficients in the log-spectral domain, therefore it is reasonable to employ diagonal covariance matrices for the GMMs in representing the models. The movement from a full covariance matrix needed for the log-spectral domain to

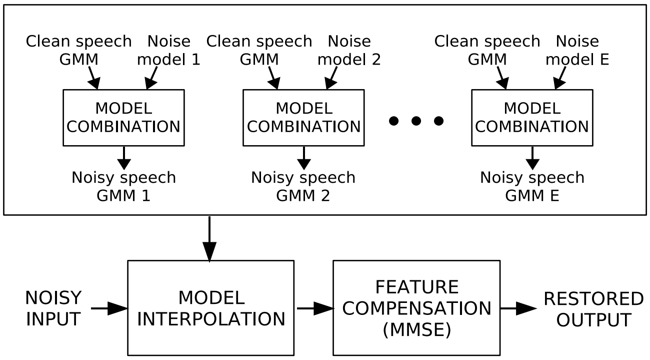

Fig. 2. PCGMM-based method employing the interpolation of multiple models.

a diagonal covariance matrix in the cepstral domain has a major reduction in both computational costs and input data requirements for more accurate model estimation. These advantages are demonstrated in the following sections.

4. PCGMM-based feature compensation employing multiple environmental models

In the proposed PCGMM-based method, model adaptation can be applied in order to address the time-varying background noise. In such a framework, the noise model is updated during silence periods via adaptation followed by combination of models, which again more accurately reflects the true noise for the GMM of the noisy speech. Such a framework however, requires considerable computational resources due to the conversion between the linear spectrum, log-spectrum and cepstral domain. Therefore, applying a model adaptation technique for the noise model may not be appropriate for small resource systems such as PDAs, navigation devices and other mobile systems. In this section, we consider the PCGMM-based method that employs a combination of environmental models for low resource based ASR applications. Utilizing multiple models estimated off-line can be effective for compensating input features adaptively under time-varying noisy conditions and eliminating the need for online model combination. Fig. 2 shows the proposed flow diagram which will be employed in the following section.

4.1. Interpolation of multiple models

Feature compensation with a single noise-corrupted speech model assumes that the recognition environment is known, and therefore employs a single previously trained noise model. However, this may not be possible in actual ASR environments, because noise conditions typically change over time. In a multiple-model method, the *a posteriori* probability of each possible environment is estimated over the incoming noisy speech. Utilizing multiple models which reflect the mixing of noisy environments represents a solution for time-varying situations.

In our work, the feature reconstruction procedure is modified using a frame-by-frame formulation for real-time processing by defining the sequential posterior probability of the environment (Kim et al., 2004). Given the incoming noisy speech feature vectors $\mathbf{Y}_t = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t]^{\mathrm{T}}$, the sequential posterior probability of a specific environment GMM G_i among E models over the input speech feature $\mathbf{Y}_t$ can be re-written as,

$$p(G_i|\mathbf{Y}_t) = \frac{P(G_i)p(\mathbf{Y}_{t-1}|G_i)p(\mathbf{y}_t|G_i)}{\sum_{e=1}^{E} P(G_e)p(\mathbf{Y}_{t-1}|G_e)p(\mathbf{y}_t|G_e)}, \tag{15}$$

where $p(\mathbf{Y}_{t-1}|G_i) = p(\mathbf{Y}_{t-2}|G_i)p(\mathbf{y}_{t-1}|G_i) = \prod_{\tau=1}^{t-1} p(\mathbf{y}_\tau|G_i)$ and $P(G_i)$ is a prior probability of each environment i represented as a GMM. Based on Eq. (14), the clean feature at frame t is reconstructed using the interpolated compensating terms as follows:

$$\hat{\mathbf{x}}_{t,\mathrm{MMSE}} \cong \mathbf{y}_t - \sum_{e=1}^{E} p(G_e|\mathbf{Y}_t) \sum_{k=1}^{K} \mathbf{r}_{e,k} p(k|G_e, \mathbf{y}_t), \tag{16}$$

where $\mathbf{r}_{e,k}$ is a constant bias term from the kth Gaussian component of the eth environment model (G_e) and $p(k|G_e, \mathbf{y}_t)$ is the posterior probability calculated from Eq. (5) for environment G_e.

When the background noise is from an environment where the number of unique types is finite, such as for in-vehicle condition (e.g., engine noise, wind noise, turn signal noise, wiper blade noise, etc. (Akbacak and Hansen, 2007)), the multiple-model method is more effective than adaptation techniques or online estimation of noise components in terms of computational complexity. In time-varying scenarios, it is also possible to employ Environmental Sniffing to detect, track, and characterize the noise types (Akbacak and Hansen, 2007). If a clean GMM is considered as one of multiple models, the performance of the recognition system can be maintained under high Signal-to-Noise Ratio (SNR) conditions. In addition, the interpolation of the clean and noise models effectively results in adaptation in time-varying or unknown SNR conditions for a particular background noise.

4.2. Special case: single noise and clean speech conditions

In particular, if a system is operating in either a clean or a single noise degraded condition, only two posterior probabilities of the clean and noisy conditions (e.g., $p(G_{\mathrm{clean}}|\mathbf{Y}_t)$ and $p(G_{\mathrm{noisy}}|\mathbf{Y}_t)$) are necessary for the multiple-model interpolation method. This represents a special case of the PCGMM multiple model approach from Section 4.1. The two posterior probabilities are re-written from Eq. (16) as follows:

$$p(G_{\mathrm{noisy}}|\mathbf{Y}_t) = \frac{P(G_{\mathrm{noisy}})p(\mathbf{Y}_{t-1}|G_{\mathrm{noisy}})p(\mathbf{y}_t|G_{\mathrm{noisy}})}{P(G_{\mathrm{noisy}})p(\mathbf{Y}_{t-1}|G_{\mathrm{noisy}})p(\mathbf{y}_t|G_{\mathrm{noisy}}) + P(G_{\mathrm{clean}})p(\mathbf{Y}_{t-1}|G_{\mathrm{clean}})p(\mathbf{y}_t|G_{\mathrm{clean}})}, \tag{17}$$

$$p(G_{\mathrm{clean}}|\mathbf{Y}_t) = 1.0 - p(G_{\mathrm{noisy}}|\mathbf{Y}_t), \tag{18}$$

where $p(\mathbf{y}_t|G_{\mathrm{noisy}}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k})$ and $p(\mathbf{y}_t|G_{\mathrm{clean}}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k})$. The summation of their prior probabilities is unity, that is, $P(G_{\mathrm{clean}}) + P(G_{\mathrm{noisy}}) = 1.0$. Therefore, if the *a priori* probabilities are assumed to be equal, Eq. (17) can be simplified as follows:

$$p(G_{\mathrm{noisy}}|\mathbf{Y}_t) = \frac{p(\mathbf{Y}_{t-1}|G_{\mathrm{noisy}})p(\mathbf{y}_t|G_{\mathrm{noisy}})}{p(\mathbf{Y}_{t-1}|G_{\mathrm{noisy}})p(\mathbf{y}_t|G_{\mathrm{noisy}}) + p(\mathbf{Y}_{t-1}|G_{\mathrm{clean}})p(\mathbf{y}_t|G_{\mathrm{clean}})}. \tag{19}$$

Finally, the estimated clean feature is obtained by the following equation:

$$\hat{\mathbf{x}}_{t,\mathrm{MMSE}} \cong \mathbf{y}_t - p(G_{\mathrm{noisy}}|\mathbf{Y}_t)\sum_{k=1}^{K}\mathbf{r}_k p(k|\mathbf{y}_t), \tag{20}$$

where $\mathbf{r}_k = \boldsymbol{\mu}_{\mathbf{y},k} - \boldsymbol{\mu}_{\mathbf{x},k}$. From Eq. (20), when there is a bi-modal clean and noisy speech sequence, for the clean sections compensation is no longer required, since the clean speech model does not contain the constant bias term. Therefore, the change of environment between clean and noisy speech conditions can be addressed by estimating the sequential posterior probability of the noisy environment without explicit detection of the condition. The interpolation between the clean speech model and noisy speech model brings in the adaptation effect under unknown SNR conditions for the specific noisy environment.

5. Computational reduction via sharing components

The amount of computation for model-based feature compensation depends primarily on the number of Gaussian components to be computed. Consequently, the computational expense increases in proportion to the number of multiple models employed for the model interpolation method described in Section 4. However, more accurate modeling for noisy conditions requires a larger number of GMMs with sufficient sized pdfs. In this section, we describe a technique of sharing the statistically similar components among the multiple environment models in an effort to reduce the computational complexity.

In the proposed method, the Gaussian components which are statistically similar to each other are selected and the common components for sharing are generated through a combining step of the similar components (Kim et al., 2004). The Kullback–Leibler distance is used to represent the separation between multi-component GMMs. The procedure of selecting the similar components is presented as follows in pseudo code, where $\mathbf{D}$ is the set of distances between Gaussian components, and $\mathbf{C}_S$ is the set of shared Gaussian components:

- **Step 0:** $\mathbf{D} = \{d_1, d_2, \ldots, d_K\}$, $\mathbf{C}_S = \emptyset$

$$d_k = \sum_{e=2}^{E} kl\_\mathrm{dist}(g_{1,k}, g_{e,k}), \quad 1 \leqslant k \leqslant K. \tag{21}$$

- **Step 1:** $\widehat{k} = \arg\ \min_k d_k \in \mathbf{D}$.
- **Step 2:** $\mathbf{C}_S = \mathbf{C}_S \cup \{\widehat{k}\}, \mathbf{D} = \mathbf{D} - \{d_{\widehat{k}}\}$.
- **Step 3:** if $N(\mathbf{C}_S) = K_S$, then stop, else go back to **Step 1**.

In the steps, d_k is the sum of Kullback–Leibler distances of the kth Gaussian component of each environmental model $g_{e,k}$ from the kth Gaussian component of the first environment $g_{1,k}$, and $N(\cdot)$ denotes the number of resulting shared elements. The first environmental model plays the pivot role in computing the distance to the Gaussians in the models. The order of the environments from *1*st to *E*th can be arbitrarily determined. Each environmental model is generated by the parallel model combination using an identical clean speech model as discussed in Section 4, so each kth Gaussian component of the environmental models is transformed from same kth component of the clean speech GMM (i.e., there is a direct pdf alignment between the environmental models). Therefore, all environmental models have the same K size GMM and it is reasonable that the Gaussian component distance across environmental models is calculated at each index k. Finally, the Gaussian search process is halted when the combined Gaussian set $\mathbf{C}_S$ reaches the desired K_S number of Gaussian components, which are now tagged as similar pdfs across the noisy speech models. The parameters of the merged Gaussian components which are shared are computed as follows:

$$\boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}} = \frac{1}{E}\sum_{e=1}^{E}\boldsymbol{\mu}_{\mathbf{y},e,k}, \quad k \in \mathbf{C}_S, \tag{22}$$

$$\boldsymbol{\Sigma}_{\mathbf{y},k}^{\{S\}} = \frac{1}{E}\sum_{e=1}^{E}\left(\boldsymbol{\Sigma}_{\mathbf{y},e,k} + (\boldsymbol{\mu}_{\mathbf{y},e,k} - \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}})(\boldsymbol{\mu}_{\mathbf{y},e,k} - \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}})^{\mathrm{T}}\right), \quad k \in \mathbf{C}_S. \tag{23}$$

The likelihood functions which contain the unique Gaussian components included in set $\mathbf{C}_S$ are replaced by the merged Gaussian components,

$$p(\mathbf{y}|e,k) = \begin{cases} p(\mathbf{y};\boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}},\boldsymbol{\Sigma}_{\mathbf{y},k}^{\{S\}}), & \text{if } k \in \mathbf{C}_S, \\ p(\mathbf{y};\boldsymbol{\mu}_{\mathbf{y},e,k},\boldsymbol{\Sigma}_{\mathbf{y},e,k}), & \text{otherwise.} \end{cases} \tag{24}$$

The constant bias terms used for feature reconstruction in Eq. (16) are also shared if their indices are included in set $\mathbf{C}_S$,

$$\mathbf{r}_{e,k} = \begin{cases} \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}} - \boldsymbol{\mu}_{\mathbf{x},k}, & \text{if } k \in \mathbf{C}_S, \\ \boldsymbol{\mu}_{\mathbf{y},e,k} - \boldsymbol{\mu}_{\mathbf{x},k}, & \text{otherwise.} \end{cases} \tag{25}$$

The computations over the $E \times K$ number of Gaussian likelihood functions can be reduced to $K_S + E(K - K_S)$, leading to a computational reduction by as much as $(E-1)K_S$ via sharing the components. If too many components are shared, performance degradation can result and therefore the number K_S must be selected to balance computational savings versus system performance. Fig. 3 illustrates the concept of the mixture sharing technique. We

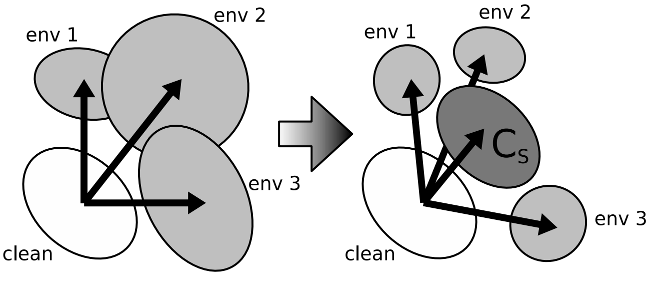

Fig. 3. Illustration of PDF mixture sharing.

should also note that *Environmental Sniffing* (Akbacak and Hansen, 2007) could be used to track and detect the complexity of the time-varying noise, which in turn could be used to help select the K_S degree of mixture sharing.

6. Experiments and results

6.1. Experimental conditions and baseline performance evaluation

The Aurora2 evaluation framework from the European Language Resources Association (ELRA) was employed to evaluate system performance (Hirsch and Pearce, 2000). The evaluation task is connected English-digits consisting of 11 words. Each whole word is represented by a continuous density HMM with 16-states and 3-mixtures per state. In addition to the digits, two silence models (i.e., normal silence and short pause) are used.

The feature extraction algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed for the experiments (ETSI, 2000). An analysis window of 25 ms duration is used with a 10 ms skip rate for 8-kHz speech data. The computed magnitude spectrum is passed through a Mel-scaled filter-bank and 23 Mel-filter-bank outputs are transformed to 13 cepstral coefficients. The 0th cepstral coefficient was used instead of the log energy, for the sake of convenience in model combination implementation. After extracting the 13th order cepstrum, the first and second order time derivatives are included during the decoding procedure (a total of the 39th order feature vector).

According to the *Clean-condition Training and Multi-condition Testing* of Aurora2, the HMM parameters were estimated using 8840 clean speech training samples and performance was evaluated with respect to each noise condition for SetA (Subway, Babble, Car and Exhibition), SetB (Restaurant, Street, Airport and Station), and SetC (Subway MIRS and Street MIRS). In SetC, the channel distortion which simulates the telecommunication terminal is also included together with the additive background noise. Each testing set consists of 1001 samples at seven different SNRs. The recognition performance cited here in the tables and figures indicate word accuracy rate, and the average value in each table was calculated based on the standard method outlined in Aurora2 (Hirsch and Pearce, 2000).

The performance of the baseline system (no compensation) is examined with comparison to several existing preprocessing algorithms in terms of environmental robustness for speech recognition. Spectral Subtraction (SS) and Cepstral Mean Normalization (CMN) were selected as the conventional algorithms. They represent the most commonly used techniques for additive noise suppression and removal of channel distortion, respectively. In spectral subtraction, the subtraction factor and flooring factor are set at 4.0 and 0.2, respectively, and background noise is estimated using the minimum statistics method with a time delay of approximately 250 ms (Martin, 1994). For cepstral mean normalization, the average value of the cepstrum over the current input utterance was subtracted from each frame. As one of conventional model adaptation methods, PMC (Gales and Young, 1996) was examined here.[1] In the PMC method, the model combination procedure is applied to the HMM speech recognizer, while the combination procedure is applied to the GMM for feature compensation in our proposed method. AFE (Advanced Front-End) algorithm suggested by ETSI was also evaluated as one of state-of-the-art methods, which contains an iterative Wiener filter and cepstral histogram equalization (ETSI, 2002). Tables 1 and 2 demonstrate performance of the baseline system and existing algorithms. From these results, we see that the combination of SS and CMN result in better performance than either method individually, and the PMC and AFE methods showed significant improvements compared to baseline and other methods.

Table 1
The recognition performance of the baseline system and conventional methods on Aurora2 test sets (word accuracy, %)

	SetA	SetB	SetC	Average
Baseline	58.56	56.67	66.16	59.32
SS	66.08	62.07	75.91	66.44
CMN	61.65	66.76	62.30	63.82
SS + CMN	73.65	77.00	74.84	75.23
PMC	81.04	81.45	76.86	80.37
AFE	85.77	84.40	84.60	84.99

6.2. Performance evaluation of the PCGMM-based method

The performance of the proposed PCGMM-based scheme is compared with other model-based feature compensation methods using identical conditions to the baseline test in Section 6.1. The GMM of the clean speech was estimated using clean speech samples identical to those used for training the HMM. The clean speech model consists of 128 Gaussian components with diagonal covariance matrices. The noise models used for model combination have a single Gaussian model and were obtained by off-line-training. The noise signals for training the model were obtained from the noise samples of Aurora2. The single Gaussian noise model was trained for each noise type and SNR condition. The speech and noise models were obtained both in the cepstral domain and in the log-spectral domain for different model-based schemes. The model-based feature compensation methods considered for comparison are as follows:

- **PCGMM**: PCGMM-based feature compensation method using model combination of the clean speech model and prior noise model trained off-line.

[1] Here, the model combination is applied to the static and delta cepstral coefficients. The mismatch function for delta–delta cepstrum is not available, because the delta–delta coefficients are obtained by a linear regression in the ETSI standard.

Table 2
The recognition performance of the baseline system and conventional methods over various SNR conditions of Aurora2 (word accuracy, %)

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB	Average
Baseline	98.82	95.39	87.33	65.76	33.81	14.35	8.16	59.32
SS	98.71	95.79	90.66	76.77	49.76	19.22	5.13	66.44
CMN	98.93	96.58	90.81	71.62	38.98	21.13	11.74	63.82
SS + CMN	98.91	97.02	94.32	86.63	64.64	32.52	15.20	75.23
PMC	98.82	97.17	95.30	90.09	74.94	44.33	19.88	80.37
AFE	99.11	97.55	95.93	91.22	81.41	58.81	27.83	84.99

- **PCGMMm**: the mean of noise model is updated with the sample mean of silence interval of each test utterance for PCGMM-based feature compensation. Approximately 200 ms duration of the silence is assumed to exist prior to the beginning of speech in every test utterance (i.e., no silence detection used). While the silence detection is important for performance assessment in real environments, the practical trade-off in WER based on speech/silence (VAD, SAD) is suggested for future work.
- **PCGMMmv**: both the mean and variance of the noise model are updated using the samples of silence duration of each utterance for PCGMM.
- **FCLS1**(Feature Compensation in the Log-Spectral domain): model combination-based feature compensation method in the log-spectral domain (Sasou et al., 2003). The means of noise-corrupted speech GMM are estimated by combining the means of the clean speech model and sample mean of silence duration using the *log-add* method. The variances of the noisy speech model are replaced with those of the clean speech model.
- **FCLS2**: model combination-based feature compensation in the log-spectral domain (Segura et al., 2001). The means and variances of the noise-corrupted speech GMM are estimated using the *log-normal approximation* method. The mean of noise model for combination is updated using the sample mean of the silence duration of the test utterances and the variance of the prior noise model.
- **VTS**(Vector Taylor Series) algorithm: feature compensation algorithm in the log-spectral domain. The noisy speech model is adaptively estimated using the EM algorithm over each test utterance (Moreno, 1996).

Table 3
The recognition performance of PCGMM-based methods and other model-based methods on Aurora2 test sets (word accuracy, %)

	SetA	SetB	SetC	Average
PCGMM	84.29	82.34	72.18	81.09
PCGMMm	**85.48**	**84.51**	81.20	**84.24**
PCGMMmv	79.44	78.91	**82.30**	79.80
FCLS1	78.90	78.64	75.64	78.14
FCLS2	83.52	84.01	76.52	82.32
VTS	75.80	77.53	76.95	76.72

As presented in Tables 3 and 4, the proposed PCGMM-based feature compensation method is effective in noisy conditions and superior performance of the PCGMM method is demonstrated compared to spectral subtraction combined with CMN and the PMC method. The results prove that the model combination used for the estimation of noisy speech GMM is effective in representing the noise corruption process. Absolute average improvements of 24.92% over baseline, and 3.15% over the basic PCGMM in word accuracy were obtained through updating the mean of the noise model (PCGMMm). This demonstrates that obtaining the sample mean from the silence interval appropriately reflects the change of noise at each utterance. However, updating the variance of the noise model resulted in a decrease in performance (PCGMMm vs. PCGMMmv). It is believed that a silence duration of approximately 200 ms was not sufficient to reliably estimate the noise variance. The comparison to other model-based methods demonstrates that the PCGMM-based compensation method in the cepstral domain is superior. In particular, FCLS2 is applied in the log-spectral domain while PCGMMm is applied in the cepstral domain. Therefore, it is suggested that the decrease in performance of FCLS2 is due to the diagonal matrices for the GMM variances in the speech distribution. The speech model in the log-spectral domain is more reliable when it has full covariance matrices, because the log-spectral coefficients are

Table 4
The recognition performance of PCGMM methods and other model-based methods over various SNR conditions of Aurora2 (word accuracy, %)

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB	Average
PCGMM	98.82	97.42	95.30	89.38	75.23	48.10	21.24	81.09
PCGMMm	98.81	**97.79**	**96.31**	**92.01**	**80.22**	**54.85**	**25.09**	**84.24**
PCGMMmv	98.79	97.40	95.40	89.83	74.19	42.17	9.78	79.80
FCLS1	98.81	97.08	94.12	86.77	69.91	42.86	19.07	78.14
FCLS2	**98.84**	97.25	95.40	90.55	77.72	50.65	22.88	82.32
VTS	98.64	96.76	94.01	86.84	69.06	36.95	15.70	76.72

Table 5
The recognition performance of PCGMM-based methods and other model-based methods combined with SS and CMN on Aurora2 test sets (word accuracy, %)

	SetA	SetB	SetC	Average
PCGMMm + SS	85.70	84.28	84.61	84.91
PCGMMm + SS + CMN	**87.21**	86.03	**87.18**	**86.73**
FCLS2 + SS + CMN	85.71	**86.29**	80.47	84.89
VTS + SS + CMN	81.06	83.75	83.48	82.62

more highly correlated with each other relative to the cepstral coefficients. The proposed method in the cepstral domain requires less data and is also more efficient than other log-spectral domain approaches due to feature dimension reduction from 23 to 13.

Tables 5, 6, and Fig. 4 present the performance of the PCGMM-based feature compensation method and other model-based methods when combined with spectral subtraction and CMN. Spectral subtraction generally increases the SNR of the noisy speech, and enhancing SNR prior to feature compensation results in more accurate discriminating posterior probabilities for reconstruction among the Gaussian components. In model combination for the PCGMM and FCLS approaches, convolutional noise such as channel distortion was not considered. Therefore, combining CMN is expected to improve performance by suppressing the channel variation across the speakers. In comparison to PCGMMm from Tables 3 and 4, combining with spectral subtraction results in a 0.67% increase in word accuracy, and combining with spectral subtraction (SS) and CMN increases the performance by 2.49%. In addition, also in other GMM-based feature compensation methods (FCLS2, VTS), combining spectral subtraction and CMN was helpful for increasing overall performance. The results demonstrate that the PCGMM-based method is superior to other model-based methods in isolation and in combination with spectral subtraction and CMN. It is also encouraging that the combined PCGMMm + SS + CMN outperforms the AFE across a wide range of SNR levels for all AURORA2 noise types in Fig. 4 (i.e., average relative WER reduction from AFE: 11.59%). Note that AFE showed the best performance when it was used in isolation without either SS or CMN.

6.3. Performance evaluation of multiple model approaches

Using the same experimental setup from Section 6.2, performance evaluation of the proposed multi-model schemes for feature compensation was also conducted. In the interpolation of multi-model PCGMM method, three different SNR-dependent noisy speech GMMs were generated using the model combination method, which are 17 dB, 7 dB, and −2 dB SNR for each noise condition. While testing a particular noise condition, different collections of noise models of the same set (A, B, and C) were used for multi-model interpolation. For example, when the test utterances of SetA were evaluated, three different SNR models in SetA (Subway, Babble, Car, and Exhibition) were employed for model interpolation. In considering the clean speech model as one environment, the number of the multiple environmental models are 13, 13, and 7 for SetA, SetB, and SetC respectively. For comparison, performance in the following combinations were examined,

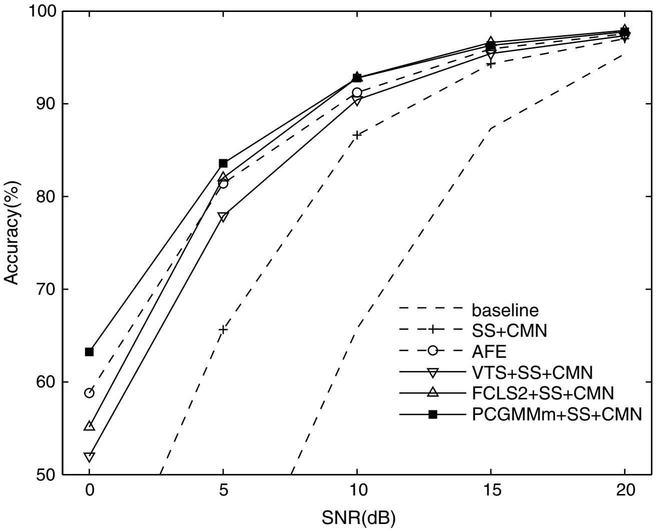

Fig. 4. The performance comparison of the PCGMM-based method and other methods combined with SS and CMN over various SNR conditions of Aurora2 (word accuracy, %).

- **IM-PCGMM**: Interpolation of Multiple Models for PCGMM-based feature compensation.
- **IM-PCGMM + SS**: IM-PCGMM combined with Spectral Subtraction.
- **IM-PCGMM + SS + CMN**: IM-PCGMM combined with Spectral Subtraction and Cepstral Mean Normalization.
- **IM-PCGMM32 + SS + CMN**: IM-PCGMM sharing 32 Gaussian components combined with Spectral Subtraction and CMN.

Table 6
The recognition performance of PCGMM-based methods and other model-based methods over various SNR conditions of Aurora2 (word accuracy, %)

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB	Average
PCGMMm + SS	98.73	97.31	95.54	91.43	81.03	59.35	29.28	84.91
PCGMMm + SS + CMN	98.87	97.76	96.30	92.77	**83.58**	**63.24**	**31.43**	**86.73**
FCLS2 + SS + CMN	**98.89**	**97.91**	**96.61**	**92.80**	82.02	55.14	23.46	84.89
VTS + SS + CMN	98.86	96.34	95.43	90.44	77.91	51.99	21.79	82.62

Table 7
The recognition performance of PCGMM-based methods using the interpolation of multiple models on Aurora2 test sets (word accuracy, %)

	SetA	SetB	SetC	Average
IM-PCGMM	85.13	83.49	70.97	81.64
IM-PCGMM + SS	85.76	83.55	80.84	83.89
IM-PCGMM + SS + CMN	87.17	85.49	85.14	86.09

- **IM-PCGMM64 + SS + CMN**: IM-PCGMM sharing 64 Gaussian components combined with Spectral Subtraction and CMN.

As presented in Tables 7 and 8, we see that PCGMM-based feature compensation schemes with the interpolation method of multiple models are effective across a range of noisy conditions, with superior performance over existing conventional algorithms. The PCGMM-based feature compensation with interpolated models (IM-PCGMM) presents similar (or even better) performance to the SNR-matched single model approach (PCGMM) which is shown in Tables 3 and 4. This proves that interpolation of multiple models is very effective for compensating the feature adaptively under blind noisy environments and changing SNR conditions in every utterance. A significant improvement was obtained by combining the IM-PCGMM method with spectral subtraction. This demonstrates that the proposed multi-model scheme is suitable for unknown SNR situations resulting after spectral subtraction. The IM-PCGMM + SS + CMN still shows better performance compared to the AFE resulting in a 7.33% average relative WER reduction compared to the AFE. Fig. 5 illustrates that the performance of the proposed multi-model approaches are comparable to noisy condition-matched single model approaches.

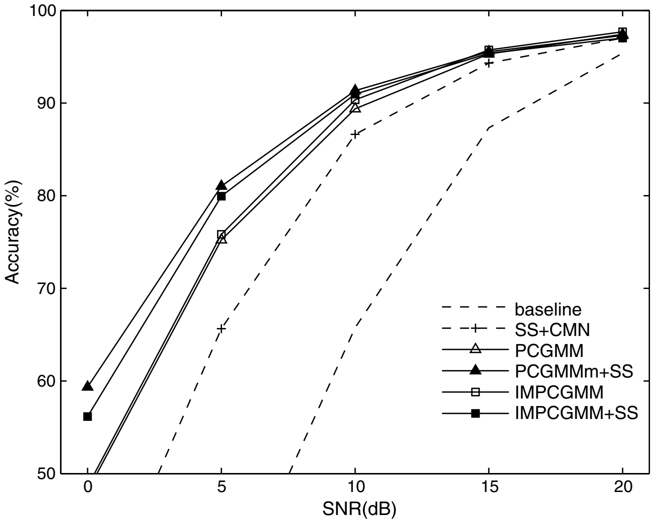

Fig. 5. The performance comparison of the PCGMM-based method using the interpolation of multiple models over various SNR conditions of Aurora2 (word accuracy, %).

Table 9
The recognition performance of multi-model PCGMM-based methods with mixture sharing on Aurora2 test sets (word accuracy, %)

	SetA	SetB	SetC	Average
IM-PCGMM + SS + CMN	87.17	85.49	85.14	86.09
IM-PCGMM32 + SS + CMN	86.46	85.11	84.44	85.52
IM-PCGMM64 + SS + CMN	85.57	84.41	83.45	84.68

Tables 9 and 10 present the performance of the IM-PCGMM-based method employing the mixture sharing technique described in Section 5. Since the combination with spectral subtraction and CMN shows a significant improvement (see Tables 7 and 8), the performance of the mixture sharing method was also considered in combination. From these results, the IM-PCGMM with the mixture sharing method (IM-PCGMM32, IM-PCGMM64) demonstrates lower performance compared to the non-sharing case. From the experiments, however, it is clear that mixture sharing is useful for reducing the computational complexity while holding the original performance at reasonable levels.

In order to investigate the relationship between performance and computational expense brought by mixture sharing, the relative WER and number of Gaussian components to be computed are summarized in Table 11. The "Difference" in the second column is the performance difference in terms of relative WER compared to the non-sharing case. The numbers in the third column ("# of Gaussian") are the number of Gaussian components to be computed for IM-PCGMM processing. In the non-sharing cases, to calculate the Gaussian probability requires 1664 (=128 × 13) components for SetA and the same number for SetB, which have 13 different noise-corrupted models. For SetC, which has seven different environment models, 896 components are needed, and therefore the

Table 8
The recognition performance of PCGMM-based methods using the interpolation of multiple models over various SNR conditions of Aurora2 (word accuracy, %)

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB	Average
IM-PCGMM	98.82	97.70	95.74	90.36	75.82	48.60	19.71	81.64
IM-PCGMM + SS	98.71	97.02	95.37	90.96	79.95	56.16	25.89	83.89
IM-PCGMM + SS + CMN	98.91	97.51	96.18	92.59	83.17	61.02	30.03	86.09

Table 10
The recognition performance of multi-model PCGMM-based methods with mixture sharing over various SNR conditions of Aurora2 (word accuracy, %)

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB	Average
IM-PCGMM + SS + CMN	98.91	97.51	96.18	92.59	83.17	61.02	30.03	86.09
IM-PCGMM32 + SS + CMN	98.91	97.23	96.00	92.29	82.67	59.40	28.79	85.52
IM-PCGMM64 + SS + CMN	98.91	97.07	95.65	91.87	81.56	57.27	27.14	84.68

Table 11
The relationship between performance and reduction in the number of Gaussian components to be computed on Aurora2 test sets

	Relative WER		Computational complexity	
	Word accuracy (%)	Difference (%)	# of Gaussian	Difference (%)
IM-PCGMM + SS + CMN	ⓐ 86.09	–	ⓓ 1510	–
IM-PCGMM32 + SS + CMN	ⓑ 85.52	−4.10 (ⓐ → ⓑ)	ⓔ 1165	22.88 (ⓓ → ⓔ)
IM-PCGMM64 + SS + CMN	ⓒ 84.68	−10.14 (ⓐ → ⓒ)	ⓕ 819	45.76 (ⓓ → ⓕ)

average number of Gaussian components to be computed becomes 1510[2] (=0.4 × 1664 + 0.4 × 1664 + 0.2 × 896) considering the proportion of the amount of test samples in Aurora2. The numerical values in the fourth column ("Difference") are the percentage of reduction in the number of Gaussian components to be computed compared to the total number of components. In the case of 32-component sharing, a 22.88% reduction in computation complexity was obtained with a 4.10% relative increase in overall WER. When 64 components are shared, a 45.76% computational reduction was achieved with a 10.14% relative increase in overall WER. As discussed in Section 5, performance decreases when the number of shared components increases. However, the experimental results demonstrate that a reasonable selection of the number of shared components will result in a significant reduction in computational complexity with an acceptable change in overall performance. This can be helpful for small footprint size mobile devices with limited storage and computational resources.

6.4. Performance in real car-driving conditions

In order to verify the effectiveness of the proposed multi-model approach in practical situations, recognition testing was accomplished on a speech corpus collected under real car-driving conditions. A number of in-vehicle corpora are available including CU-Move (Hansen et al., 2004), UTDrive (Angkititrakul et al., 2007), CIAIR (Kawaguchi et al., 2004), and SITEC (http://www.sitec.or.kr). Here, we use the Car01 and CarNoise01 corpus released by the Speech Information Technology and Industry Promotion Center (SITEC). Car01 contains Korean speech utterances recorded in a car-driving at a speed of 80 km/h. CarNoise01 contains noise samples recorded in various driving situations.

For recognition testing, a 548 vocabulary set was chosen in Car01 consisting of control command words in the vehicle. Fig. 6 presents the locations of microphones used for recording the Car01 database. A total of 4384 utterances recorded via a head-set microphone (channel 1) were used for clean HMM training and 1096 utterances for noisy condition testing which were recorded via a directional microphone located at the center of driver's sun visor (channel 4). Table 12 presents the performance of the baseline system and conventional methods with Car01 data. The performance of the PCGMM-based feature compensation methods are presented in Table 13. PCGMM denotes the PCGMM-based feature compensation method with a single model and the noise model for model combination

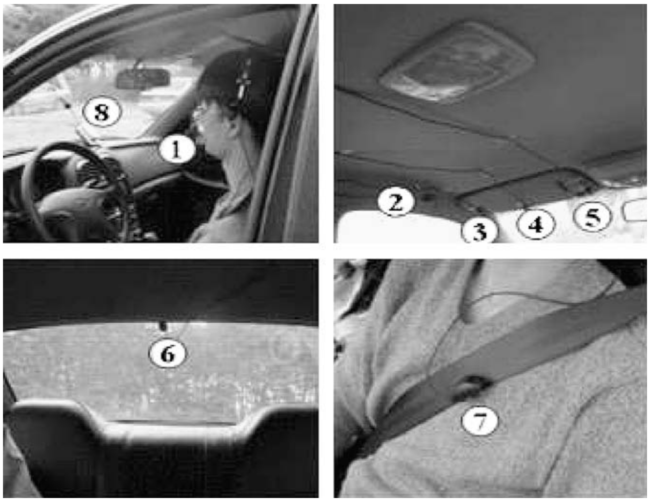

Fig. 6. The locations of microphones used for collecting the speech database Car01 under real car-driving condition.

Table 12
The recognition performance of the baseline system and conventional methods in the real car-driving condition Car01 database (word accuracy, %)

Clean (ch1)	Noisy (ch4)	SS (ch4)	SS+CMN (ch4)
94.16	58.76	82.94	88.96

[2] This value is a fraction, but we take the largest integer to represent the number of Gaussian components.

Table 13
The recognition performance of the PCGMM-based methods using single and multiple models in the real car-driving condition, channel 4 microphone of Car01 database (word accuracy, %)

PCGMM	IM-PCGMM	IM-PCGMM + SS + CMN	IM-PCGMM64 + SS + CMN
88.96	88.96	91.33	91.24

Table 14
The relationship between performance and reduction in the number of Gaussian components to be computed in the real car-driving conditions, channel 4 microphone of Car01 database

	Relative WER		Computational complexity	
	Word accuracy (%)	Difference (%)	# of Gaussian	Difference (%)
IM-PCGMM + SS + CMN	ⓐ 91.33	–	ⓒ 512	–
IM-PCGMM64 + SS + CMN	ⓑ 91.24	−1.04 (ⓐ → ⓑ)	ⓓ 320	37.50 (ⓒ → ⓓ)

was estimated from the noise samples in CarNoise01 recorded while driving at a speed of 80 km/h. For PCGMM with multi-model interpolation (IM-PCGMM), three kinds of noise models were used, which were estimated from the noise samples of 50 km/h, 80 km/h and 100 km/h. From the table, we see that the proposed multi-model scheme (IM-PCGMM) showed comparable performance to conventional methods (SS or SS + CMN) and condition-matched single model method (PCGMM) in real-life environments. The comparable performance of the spectral subtraction method (SS + CMN) to the proposed methods (PCGMM, IM-PCGMM) here is considered due to the background noise characteristics of Car01 database which has a relatively high SNR (i.e., 7–8 dB) and is highly stationary with low-frequency content during each test utterance. The multi-model scheme combined with SS and CMN (IM-PCGMM + SS + CMN) outperforms all other methods/combinations. This illustrates that multiple noisy speech models at different speeds is effective in reflecting the range of background noise at unknown speeds for in-vehicle conditions. The results in Table 14 show that the proposed mixture sharing technique produces significant reduction in computational complexity (e.g., 37.50%) while maintaining recognition performance (e.g., only a 1.04% reduction). The performance “Difference” in the second column was calculated in terms of relative WER compared to the non-sharing case (IM-PCGMM + SS + CMN).

7. Conclusions

In this study, a feature compensation algorithm employing a combination of GMMs operating in the cepstral domain was developed. The proposed scheme eliminates the need for a prior noisy speech database in the training procedure, by applying model combination to the estimation of the noisy speech model across a range of SNRs and noise types. The proposed scheme has several advantages including computational reduction and more accurate modeling, as applied in the cepstral domain. The interpolation method of multiple noise environmental models was employed to address time-varying noisy conditions. In order to reduce the computational expense due to multiple models, a sharing technique for similar noisy Gaussian speech components was also proposed. In order to evaluate the performance of the developed schemes, speech recognition experiments were performed using both simulated adverse environments (e.g., Aurora2), and actual in-vehicle conditions. The experimental results consistently demonstrated that the cepstral feature compensation method based on model combination is more effective, compared to other existing feature compensation methods. Employing multiple models proved to be effective in addressing changing noisy speech conditions comparable to the environment-matched model. The mixture sharing technique was helpful in significantly reducing computational expenses while holding recognition accuracy at an acceptable level.

Future work could consider applying this method in conjunction with *Environmental Sniffing* (Akbacak and Hansen, 2007) in order to prune a much larger library of noise environments, resulting in a more focused noise compensation scheme. For example, if there were 1000 noise GMMs, an environmental sniffer could prune this library to a sub-set of car, truck, or mobile environment of interest. The method could also be applied for other speech applications such as speaker ID, language ID, and others. Finally, it could be employed for automatic transcript generation in spoken document retrieval using speech recognition requiring sustained performance over a wide diversity of acoustic conditions (Hansen et al., 2005).

References

Acero, A., 1993. Acoustic and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publisher.

Akbacak, M., Hansen, J.H.L., 2007. Environmental sniffing: noise knowledge estimation for robust speech systems. IEEE Trans. Audio Speech Lang. Process. 15 (2), 465–477.

Angkititrakul, P., Petracca, M., Sathyanarayana, A., Hansen, J.H.L., 2007. UTDrive: driver behavior and speech interactive systems for in-vehicle environments. In: IEEE Intelligent Vehicles Symposium, pp. 566–569.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Droppo, J., Deng, L., Acero, A., 2001. Evaluation of the SPLICE algorithm on the Aurora 2 Database. In: Eurospeech2001, pp. 217–220.

Ephraim, Y., Malah, D., 1984. Speech enhancement using minimum mean square error short time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

ETSI standard document, 2000. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms. ETSI ES 201 108 v1.1.2 (2000-04).

ETSI standard document, 2002. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. ETSI ES 202 050 v1.1.1 (2002-10).

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4 (5), 352–359.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Hansen, J.H.L., 1994. Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect. IEEE Trans. Speech Audio Process. 2 (4), 598–614.

Hansen, J.H.L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Commun. 20 (2), 151–170.

Hansen, J.H.L., Arslan, L., 1995. Robust feature estimation and objective quality assessment for noisy speech recognition using the credit card corpus. IEEE Trans. Speech Audio Process. 3 (3), 169–184.

Hansen, J.H.L., Clements, M., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39 (4), 795–805.

Hansen, J.H.L., Womack, B., 1996. Feature analysis and neural network based classification of speech under stress. IEEE Trans. Speech Audio Process. 4 (4), 307–313.

Hansen, J.H.L., Zhang, X., Akbacak, M., Yapanel, U., Pellom, B., Ward, W., Angkititrakul, P., 2004. CU-Move: advances for in-vehicle speech systems for route navigation. In: Abut, Hansen, Taketa (Eds.), DSP for In-Vehicle and Mobile Systems. Springer, Chapter 2.

Hansen, J.H.L., Huang, R., Chou, B., Beadle, M., Deller Jr., J.R., Gurijala, A.R., Kurimo, M., Angkititrakul, P., 2005. SpeechFind: advances in spoken document retrieval for a national gallery of the spoken word. IEEE Trans. Speech Audio Process. 13 (5), 712–730.

Hirsch, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. ISCA ITRW ASR2000.

Kawaguchi, N., Matsubara, S., Kishida, I., Irie, Y., Murao, H., Yamaguchi, Y., Takeda, K., Itakura, F., 2004. Construction and analysis of a multi-layered in-car spoken dialogue corpus. In: Abut, Hansen, Taketa (Eds.), DSP for In-Vehicle and Mobile Systems. Springer, Chapter 1.

Kim, N.S., 2002. Feature domain compensation of nonstationary noise for robust speech recognition. Speech Commun. 37, 231–248.

Kim, W., Ahn, S., Ko, H., 2003. Feature compensation scheme based on parallel combined mixture model. In: Eurospeech2003, pp. 677–680.

Kim, W., Kwon, O., Ko, H., 2004. PCMM-based feature compensation schemes using model interpolation and mixture sharing. In: ICASSP2004, pp. 989–992.

Lee, K.F., 1989. Automatic Speech Recognition: The Development of the SPHINX system. Kluwer Academic Publisher.

Lee, C.H., Lin, C.H., Juang, B.H., 1991. Study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Trans. Signal Process. 39 (4), 806–814.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Comput. Speech Lang. 9, 171–185.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: EUSIPCO-94, pp. 1182–1185.

Morales, N., Toledano, D.T., Hansen, J.H.L., Garrido, J., Colas, J., 2006. Unsupervised class-based feature compensation for time-variable bandwidth-limited speech. In: ICASSP2006, pp. 533–536.

Moreno, P.J., 1996. Speech recognition in noisy environments. Ph.D. Thesis. Carnegie Mellon University.

Moreno, P.J., Raj, B., Stern, R.M., 1998. Data-driven environmental compensation for speech recognition: a unified approach. Speech Commun. 24 (4), 267–285.

Raj, B., Stern, R.M., 2005. Missing-feature approaches in speech recognition. IEEE Signal Process. Mag. 22 (5).

Sasou, A., Asano, F., Tanaka, T., Nakamura, S., 2003. Adaptation of acoustic model using the gain-adapted HMM decomposition method. In: Eurospeech2003, pp. 29–32.

Sasou, A., Tanaka, T., Nakamura, S., Asano, F., 2004. HMM-based feature compensation methods: an evaluation using the Aurora2. In: ICSLP2004, pp. 121–124.

Schulte-Fortlcamp, B., Brooks, B.M., Bray, W.R., 2007. SoundScape: an approach to rely on human perception and expertise in the post-modern community noise era. Acoust. Today 3 (1), 7–15.

Segura, J.C., Torre, A., Benitez, M.C., Peinado, A.M., 2001. Model-based compensation of the additive noise for continuous speech recognition: experiments using the Aurora II database and tasks. In: Eurospeech2001, pp. 221–224.

Singh, R., Stern, R.M., Raj, B., 2002. Signal and feature compensation methods for robust speech recognition. Chapter in CRC Handbook on Noise Reduction in Speech Applications. CRC Press.

Stouten, V., Van hamme, H., Wambacq, P., 2004. Joint removal of additive and convolutional noise with model-based feature enhancement. In: ICASSP2004, pp. 949–952.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: ICASSP90, pp. 845–848.

Westphal, M., Waibel, A., 2001. Model-combination-based acoustic mapping. In: ICASSP2001, pp. 221–224.

Womack, B., Hansen, J.H.L., 1996. Robust speech recognition via speaker stress classification. In: ICASSP-96, vol. 1, pp. 53–56.

Womack, B., Hansen, J.H.L., 1999. N-channel hidden Markov models for combined stressed speech classification and recognition. IEEE Trans. Speech Audio Process. 7 (6), 668–677.